Report on INEX 2012

P. Bellot T. Chappell A. Doucet S. Geva S. Gurajada J. Kamps
G. Kazai M. Koolen M. Landoni M. Marx A. Mishra V. Moriceau
J. Mothe M. Preminger G. Ramírez M. Sanderson E. Sanjuan F. Scholer
A. Schuh X. Tannier M. Theobald M. Trappett A. Trotman Q. Wang

Abstract

INEX investigates focused retrieval from structured documents by providing large test collections of structured documents, uniform evaluation measures, and a forum for organizations to compare their results. This paper reports on the INEX'12 evaluation campaign, which consisted of a five tracks: Linked Data, Relevance Feedback, Snippet Retrieval, Social Book Search, and Tweet Contextualization. INEX'12 was an exciting year for INEX in which we joined forces with CLEF and for the first time ran our workshop as part of the CLEF labs in order to facilitate knowledge transfer between the evaluation forums.

1 Introduction

Traditional IR focuses on pure text retrieval over "bags of words" but the use of structure—such as document structure, semantic metadata, entities, or genre/topical structure—is of increasing importance on the Web and in professional search. INEX has been pioneering the use of structure for focused retrieval since 2002, by providing large test collections of structured documents, uniform evaluation measures, and a forum for organizations to compare their results.

INEX'12 was an exciting year for INEX in which we joined forces with CLEF and ran our workshop as part of the CLEF labs in order to foster further collaboration and facilitate knowledge transfer between the evaluation forums. In total five research tracks were included, which studied different aspects of focused information access:

Linked Data Track investigating retrieval over a strongly structured collection of documents based on DBpedia and Wikipedia. The *Ad Hoc Search Task* has informational requests to be answered by the entities in DBpedia/Wikipedia. The *Faceted Search Task* asks for a restricted list of facets and facet-values that will optimally guide the searcher toward relevant information.

Relevance Feedback Track investigate the utility of incremental passage level relevance feedback by simulating a searcher's interaction. An unconventional evaluation track where submissions are executable computer programs rather than search results.

Snippet Retrieval Track investigate how to generate informative snippets for search results. Such snippets should provide sufficient information to allow the user to determine the relevance of each document, without needing to view the document itself.

Social Book Search Track investigating techniques to support users in searching and navigating collections of digitised or digital books, metadata and complementary social media. The *Social Book Search Task* studies the relative value of authoritative metadata and user-generated content using a test collection with data from Amazon and LibraryThing. The *Prove It Task* asks for pages confirming or refuting a factual statement, using a corpus of the full texts of 50k digitized books.

Tweet Contextualization Track investigating tweet contextualization, answering questions of the form "what is this tweet about?" with a synthetic summary of contextual information grasped from Wikipedia and evaluated by both the relevant text retrieved, and the "last point of interest."

In the rest of this paper, we discuss the aims and results of the INEX'12 tracks in relatively self-contained sections: the Linked Data track (Section 2), the Relevance Feedback track (Section 3), the Snippet Retrieval track (Section 4), the Social Books Search track (Section 5), and the Tweet Contextualization track (Section 6),

2 Linked Data Track

In this section, we will briefly discuss the INEX'12 Data Centric Track. Further details are in [7].

2.1 Aims and Tasks

The goal of the new Linked Data Track was to investigate retrieval techniques over a combination of textual and highly structured data, where rich textual contents from Wikipedia articles serve as the basis for retrieval and ranking, while additional RDF properties from DBpedia and YAGO2 carry structural information about entities and semantic relations among entities extracted from articles. Our intention in organizing this new track thus follows one of the key themes of INEX, namely to explore and investigate if and how structural information could be exploited to improve the effectiveness of ad-hoc retrieval. In addition, we were interested in how this combination of data could be used together with structured queries to help users navigate or explore large sets of results (a task that is well-known from faceted search systems), or to address Jeopardy-style natural-language clues or questions (known, for example, from recent question answering settings over linked data collections). The Linked Data Track thus aims to close the gap between IR-style keyword search and semantic-web-style reasoning techniques, with the goal to bring together different communities and to foster research at the intersection of Information Retrieval, Databases, and the Semantic Web.

For INEX 2012, we explored three different retrieval tasks: The classic *Ad Hoc Search Task* investigates informational queries to be answered mainly by the textual contents of the Wikipedia articles. The *Faceted Search Task* employs facets and facet-values obtained from the DBpedia ontology and aims to guide the searcher toward relevant information. The new *Jeopardy Task* employs natural-language Jeopardy clues which are manually translated into a semi-structured query format based on SPARQL with keyword filter conditions.

2.2 Test Collection

The core data collection of the Linked Data Track consists of Wikipedia articles and RDF properties from DBpedia 3.7 and YAGO2. Each Wikipedia article corresponds to an entity/resource in DBpedia and YAGO2. The connection between the data sets is given in the "wikipedia links en.nt" file from DBpedia. The Wikipedia articles in the collection was based on the MediaWiki-formated dump dated on July 22, 2011, which corresponds to the version 3.7 of DBpedia. To facilitate participants process the data collection, we built a fused collection of XML files by first converting each raw Wikipedia article (originally in MediaWiki markup) into a customized XML-formatted file, and then appending the RDF triples imported from both DBpedia and YAGO2 that contain the article entity as subject or object to the article's XML file. Unfortunately, due to the complexity of MediaWiki markup, the parser employed failed in parsing all articles successfully, and thus resulted in a subset of Wikipedia articles in the fused collection.

In addition, participants were explicitly encouraged to make use of more RDF data from other sources (see, for example, linkeddata.org) that go beyond "just" DBpedia and YAGO2. Any inclusion of further data sources was welcome, however, workshop submissions and follow-up research papers should explicitly mention these sources when describing their approaches. A dedicated topics set was designed, consisting of general topics, subtopics, and subsubtopics (refinements of subtopics) for the evaluation of facetted search and ad hoc search.

2.3 Results

In total, 20 ad-hoc search runs were submitted by 7 participants and 5 valid Jeopardy! runs were submitted by 2 participants. The facetted search task is still ongoing due to the late availability of the corpus. Assessment was done using the Amazon Mechanical Turk. The 30 sub-subtopics collected from the faceted search task and 50 randomly selected Jeopardy! topics were assessed, while the relevance results for the 20 old topics in INEX 2009 and 2010 were reused in evaluation. The TREC MAP metric, as well as P@5, P@10, P@20 and so on, was used to measure the performance of all ad-hoc and Jeopardy runs. For the Faceted Search Task, we will use the same evaluation metrics as that used in last year.

Over all the 100 ad hoc topics, the best performing run in terms of MAP was from Renmin University of China, and the second and third best scoring teams were University of Otago and Ecole des Mines de Saint-Etienne respectively. Their precision-recall curves, however, were quite close to each other. They all used traditional IR approaches, except that Renmin University combined the retrievals on text and RDF data. Over the 50 Jeopardy! topics, the best scoring run in terms of the mean reciprocal rank was submitted by Renmin University of China, University of Amsterdam being the second best scoring team. All the teams except the Max-Planck Institute for Informatics based their approaches only on the keywords not on the SPARQL FullText queries of the Jeopardy questions .

2.4 Outlook

The Linked Data Track, a new track in INEX 2012, was organized towards our goal to close the gap between IR-style keyword search and semantic search techniques. We believe that this track encourages further research towards applications that exploit semantic annotations over large text collections and thus facilitates the development of effective retrieval techniques for the same. The track will continue in INEX 2013 with emphasis on complex retrieval tasks such as Jeopardy! questions and a new better preprocessed data collection (based on DBpedia 3.8, with parsing errors of MediaWiki markup resolved and so on), which could ease the participations of both IR and Semantic Web researchers.

3 Relevance Feedback Track

In this section, we will briefly discuss the INEX'12 Relevance Feedback track. Further details are in [1].

3.1 Aims and Task

The track was designed to facilitate the development of search engine modules that incorporate focused relevance feedback. It covers the use of focused feedback, a relevance feedback model wherein users specify segments of the document (usually through some form of selection or highlighting tool) considered relevant to the search topic. This allows users to give more flexible feedback when only portions of the current document are relevant to their search.

Organizations participated by supplying executables that would communicate with a supplied evaluation platform through standard operating system I/O pipes. The evaluation platform would provide the search topics and, for each document provided to it by the search module, reply with relevant passages. The search module can then make use of this information to rerank the remaining documents as necessary.

After each topic has been searched, the evaluation platform uploads the document IDs returned by the search module for each topic in the form of a *trec_eval*-compatible submission. The submission is evaluated on a remote server against relevance assessments for the topics and the results are sent back to the evaluation platform.

3.2 Test Collection

The INEX Wikipedia Collection, a 50.7GB collection of 2,666,190 Wikipedia articles in XML format was used as the data collection for the track.

The search topics and assessments used were collected for the INEX 2009 and 2010 Ad Hoc tracks. 10 topics were used for the training set and 50 were used for the evaluation set.

3.3 Submissions

Queensland University of Technology made five submissions using an experimental relevance feedback mode in TopSig. The TopSig runs, apart from the baseline run which did not make use of feedback at all, simply used the feedback text as a new query and reranked the remaining documents found by the initial query each time.

The Universidad Autónoma Metropolitana made 10 submissions using Indri as a base and employing a Markov random field to rerank results with relevance feedback. The BASE-IND run consists of a run with Indri without incorporating relevance feedback while the MF and LF runs consist of the results when adding the 20 most frequent and least frequent terms respectively from the feedback to the query. The RRMRF runs are also based on Indri but employ the Markov random field for reranking.

3.4 Results

For UAM their baseline run BASE-IND obtained a MAP of 0.1015 and an R-Precision of 0.1828. Their incremental feedback run RRMRF did not improve performance. For QUT, their baseline TOPSIG-2048 scored a MAP of 0.2218 and an R-Precision of 0.2688. Their best score feedback run scored a MAP of 0.2477 and an R-Precision of 0.2812.

3.5 Outlook

We have presented the Relevance Feedback track at INEX 2012. While the number of participants has not increased since the last iteration of the track, the new approach used allows much greater flexibility in the evaluation of focused relevance feedback approaches.

4 Snippet Retrieval Track

In this section, we will briefly discuss the INEX 2012 Snippet Retrieval Track. Further details are in [6].

4.1 Aims and Task

The goal of the snippet retrieval track is to determine how best to generate informative snippets for search results. Such snippets should provide sufficient information to allow the user to determine the relevance of each document, without needing to view the document itself, allowing the user to quickly find what they are looking for.

The task was to return a ranked list of documents for the requested topic to the user, and with each document, a corresponding text snippet describing the document. Each run was to return 20 documents per topic, with a maximum of 180 characters per snippet.

4.2 Collection

The Snippet Retrieval Track uses the INEX Wikipedia collection introduced in 2009—an XML version of the English Wikipedia, based on a dump taken on 8 October 2008, and semantically annotated as described by Schenkel et al. [5]. This corpus contains 2,666,190 documents.

There were 35 topics in total—10 taken from the INEX 2010 Ad Hoc Track, and 25 created specifically for this track, with the goal being to create topics requesting more specific information than is likely to be found in the first few paragraphs of a document. Each topic contains a short content only (CO) query, a phrase title, a one line description of the search request, and a narrative with a detailed explanation of the information need, the context and motivation of the information need, and a description of what makes a document relevant or not.

4.3 Assessment and Evaluation

To determine the effectiveness of the returned snippets at their goal of allowing a user to determine the relevance of the underlying document, manual assessment is being used. In response to feedback from the previous year, both snippet-based and document-based assessment are being used.

The documents will first be assessed for relevance based on the snippets alone, as the goal is to determine the snippet's ability to provide sufficient information about the document. The documents will then be assessed for relevance based on the full document text, with evaluation based on comparing these two sets of assessments. Each topic within a submission is assigned an assessor. The assessor, after reading the details of the topic, reads through the 20 returned snippets, and judges which of the underlying documents seem relevant based on the snippets. The assessor is then presented the full text of each document, and determines whether or not the document is actually relevant. To avoid bias introduced by assessing the same topic more than once in a short period of time, and to ensure that each submission is assessed by the same assessors, the runs are shuffled in such a way that each assessment package contains one run from each topic, and one topic from each submission.

Submissions are evaluated by comparing the snippet-based relevance judgements with the document-based relevance judgements, which are treated as a ground truth. The primary evaluation metric used is the geometric mean of recall and negative recall (GM). A high value of GM requires a high value in recall and negative recall—i.e., the snippets must help the user to accurately predict both relevant and irrelevant documents. If a submission has high recall but zero negative recall (e.g. in the case that everything is judged relevant), GM will be zero. Likewise, if a submission has high negative recall but zero recall (e.g. in the case that everything is judged irrelevant), GM will be zero. Details of additional metrics used are given in [6].

4.4 Results

As of this writing, the assessment and evaluation phase is yet to begin. Once completed, results will be available in [6].

5 Social Book Search Track

In this section, we will briefly discuss the INEX'12 Social Book Search Track. Further details are in [3].

5.1 Aims and Tasks

Prompted by the availability of large collections of digitized books, the Social Book Search Track aims to promote research into techniques for supporting users in searching, navigating and reading full texts of digitized books and associated metadata. This year, the track ran two tasks: the Social Book Search task and the Prove It task:

1. The *Social Book Search* (SBS) task, framed within the scenario of a user searching a large online book catalogue for a given topic of interest, aims at exploring techniques to deal with both complex information needs of searchers—which go beyond topical relevance and can include aspects such as genre, recency, engagement, interestingness, quality and how well-written a book is—and complex information sources including user profiles and personal catalogues, and book descriptions containing both professional metadata and user-generated content.

2. The *Prove It* (PI) task aims to test focused retrieval approaches on collections of books, where users expect to be pointed directly at relevant book parts that may help to confirm or refute a factual claim;

In addition to these task, there are two related projects. The *Structure Extraction* (SE) task, running at ICDAR in 2013, aims at evaluating automatic techniques for deriving structure from OCR and building hyperlinked table of contents. The *Active Reading task* (ART) aims to explore suitable user interfaces to read, annotate, review, and summarize multiple books.

5.2 Test Collections

For the Social Book Search task a new type of test collection has been developed. Unlike traditional collections of topics and topical relevance judgements, the task is based on rich, real-world information needs from the LibraryThing discussion forums and user profiles. The collection consists of 2.8 million book descriptions from Amazon, including user reviews, and is enriched with user-generated content from LibraryThing. For the information needs we used the LT discussion forums. We selected discussion threads started by members who ask for recommendations on a certain topic, which often contain detailed descriptions of they are looking for. Associated with this request is the personal catalogue of that user, with an entry date and tags that the user associated with each book. Together, these form a rich description of the user's information need. The relevance judgements come in the form of suggestions from other LT members in the same discussion thread. We compared the suggestions with the personal catalogue of the requester to identify which of the suggestion she subsequently catalogued. These are considered the most relevant suggestions.

The PI task builds on a collection of over 50,000 digitized out-of-copyright books of different genre (e.g., history books, text books, reference works, novels and poetry) marked up in XML. The task was first run in 2010 and was kept the same for 2011 and 2012. This year the evaluation is based on 30 of the 83 topics created in 2010. Most topics consist of complex factual statements with multiple atomic facts. This creates difficulties for judging, as individual parts can be confirmed or refuted on a page without the other parts being mentioned. Therefore, we chose to split the complex statements into atomic aspects that can more easily be judged. We are currently running an experiment on Amazon Mechanical Turk to obtain judgements for the individual aspects.

5.3 Results

Four teams submitted 17 runs to the Social Book Search task and three teams submitted 12 runs to the Prove It! task. The Social Book Search task evaluation has shown that the most effective systems incorporate the full topic statement, which includes the title of the topic thread, the name of the discussion group and the full first message that elaborates on the request. However, the best system did not use any user profile information. So far, the best system is a plain full-text retrieval system.

For the Prove It! task, there are no evaluation results yet. We expect to have judgements from Mechanical Turk and evaluation results in time for the final INEX proceedings.

5.4 Outlook

Next year, we continue with the Social Book Search task to further investigate the role of user information. We also plan to enrich the relevance judgements with further judgements on the relevance of books to specific relevance aspects of the information need. For this, we plan to use either Mechanical Turk or approach the topic creators on LibraryThing to obtain more specific judgements directly from the person with the actual information need. We also plan to incorporate a large set of new topics, by asking Library and Information Science students to provide rich labels and mediated search queries that hopefully better reflect how searchers would express their needs in search queries.

This year the Prove It task has undergone some changes. Next year we also hope to enhance the task by explicitly considering the authoritativeness of a source that contains primary or secondary evidence that confirms or refutes a given factual topic statement.

6 Tweet Contextualization Track

In this section, we will briefly discuss the INEX'12 Tweet Contextualization Track. Further details are in [4].

6.1 Aims and Tasks

The use case of the Tweet Contextualization (TC) Track is the following: given a new tweet, participating systems must provide some context about the subject of a tweet, in order to help the reader to understand it. In this task, contextualizing tweets consists in answering questions of the form "what is this tweet about?" which can be answered by several sentences or by an aggregation of texts from different documents of the Wikipedia. The general process involves tweet analysis, passage and/or XML element retrieval, and construction of the context/summary.

For evaluation purposes, we require that a summary uses only elements or passages previously extracted from the document collection and does not exceed 500 words. The correctness of summaries is established exclusively based on the support passages and documents. The summaries are evaluated according to Informativeness (the way they overlap with relevant passages) and Readability (assessed by evaluators and participants).

6.2 Test Collection

Organizers provided a document collection extracted from Wikipedia, as well as 1,000 English topics made of tweets from several different accounts. The document collection has been built based on a recent dump of the English Wikipedia from November 2011. This date is anterior to all selected tweets. We removed all notes and bibliographic references that are difficult to handle and kept only non empty Wikipedia pages.

Tweets were collected by the track organizers from Twitter® Search API. They were selected among informative accounts, in order to avoid purely personal tweets that could not be contextualized. The organizers selected 63 of these tweets manually. For each of them, we checked that the document collection contained some information related to the topic of the tweet. This means that all 63 tweets had some contextualization material inside the provided collection. All 1,000 tweets were to be treated by participants, which ensured that only robust systems could achieved the task. However, only those 63 checked tweets were used for evaluation.

6.3 Measures

Tweet contextualization is evaluated on both informativeness and readability. Informativeness aims at measuring how well the summary explains the tweet or how well the summary helps a user to understand the tweet content. On the other hand, readability aims at measuring how clear and easy to understand the summary is.

The informativeness of summaries is evaluated by comparing the probability $P(t|T)$ of finding an n-gram t in the reference T to the probability $P(t|S)$ of finding this n-gram in the submitted summary S. More precisely the metric stands as:

$$LogSim(T,S) \quad = \quad \sum_{t \in T} P(t|T) \times \frac{\min(\log(1+P(t|T)), \log(1+P(t|S)))}{\max(\log(1+P(t|T)), \log(1+P(t|S)))} \qquad (1)$$

The evaluation of readability uses binary questionnaires that evaluate syntactic problems, unsolved anaphora, redundancy, and un-understandingness. As opposed to Informativeness, readability evaluation cannot be reproduced on unofficial runs.

6.4 Results

In total 33 valid runs by 13 teams from 10 countries were submitted. Only three teams among the 13 used the perl API and Indri index provided by organizers.

The release of reusable assessments to evaluate text informativeness is one of the important results of TC track. For each topic, all passages from all participants have been merged and displayed to the assessors in alphabetical order. Therefore, each passage informativeness has been evaluated independently from others, even in the same summary. Readability of 594 summaries from 18 tweets have been assessed (highly incoherent grammatical structures, unsolved anaphora, redundant passages). Each participant had to evaluate readability for a pool of abstracts on an online web interface.

Since 2011 previous edition, maximal informativeness score increased from 10% to 14% and the third quantile from 8% to 9%. Readability scores remained similar. All participants but two used language models, however informativeness of runs that only used passage retrieval is under 5%. All systems having a run among the top five in both informativeness and readability rankings used state of the art PoS parsing and summarization methods based on sentence scoring. Tweet reformulation based on general or local Latent Dirichlet Allocation (LDA) was used by two participants. Local LDA performed better, reaching 12% of informativeness with an acceptable level of readability (above 70%).

6.5 Outlook

The use case and the topic selection should remain stable in 2013 TC Track, so that 2012 topics can be used as a training set. Nevertheless, we will consider more diverse types of tweets, so that participants could better measure the impact of hashtag processing on their approaches.

7 Envoi

This complete our walk-through of the five tracks of INEX'12. The tracks cover various aspects of focused retrieval in a wide range of information retrieval tasks. This report has only touched upon the various approaches applied to these tasks, and their effectiveness. The formal proceedings of INEX'12 are being published in the Springer LNCS series [2]. This volume contains both the track overview papers, as well as the papers of the participating groups. The main result of INEX'12, however, is a great number of test collections that can be used for future experiments.

INEX'13 will seek to continue the successful collaboration with CLEF, and to organize the workshop as part of the CLEF labs in Valencia in September 2013. The Relevance Feedback track will continue as an online evaluation platform, where new runs can be uploaded and evaluated at any time. The other four tracks will continue the new set-up of INEX'12, with plans for a new corpus based on Wikipedia that aligns work in the Linked Data, Snippet Retrieval, and Tweet Contextualization tracks.

References

[1] T. Chappell and S. Geva. Overview of the INEX 2012 relevance feedback track. In Geva et al. [2].

[2] S. Geva, J. Kamps, and R. Schenkel, editors. *Focused Access to Content, Structure and Context: 11th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX'12)*, LNCS. Springer, 2013.

[3] M. Koolen, G. Kazai, J. Kamps, M. Preminger, A. Doucet, and M. Landoni. Overview of the INEX 2012 social book search track. In Geva et al. [2].

[4] E. Sanjuan, V. Moriceau, X. Tannier, P. Bellot, and J. Mothe. Overview of the INEX 2012 tweet contextualization track. In Geva et al. [2].

[5] R. Schenkel, F. M. Suchanek, and G. Kasneci. YAWN: A semantically annotated Wikipedia XML corpus. In *12. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW 2007)*, pages 277–291, 2007.

[6] M. Trappett, S. Geva, A. Trotman, F. Scholer, and M. Sanderson. Overview of the INEX 2012 snippet retrieval track. In Geva et al. [2].

[7] Q. Wang, J. Kamps, G. Ramírez, M. Marx, A. Schuth, M. Theobald, S. Gurajada, and A. Mishra. Overview of the INEX 2012 data centric track. In Geva et al. [2].